DOCUMENT RESUME

ED 451 238 TM 032 466

AUTHOR Bastick, Tony
TITLE Reliability Problems of the Datum: Solutions for Questionnaire Responses.
PUB DATE 1999-07-00
NOTE 8p.; Paper presented at the European Meeting of the Psychometric Society (11th, Luneburg, Germany, July 1999).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Questionnaires; *Reliability; *Responses
IDENTIFIERS Estimation

ABSTRACT

Questionnaires often ask for estimates, and these estimates are given with different reliabilities. It is difficult to know the different reliabilities of single estimates and to take these into account in subsequent analyses. This paper contains a practical example to show that not taking the reliability of different responses into account can lead to erroneous conclusions. A solution is suggested in which two estimates are requested that are then used as upper and lower bounds. The mean of this double estimate then acts as equivalent to, or more accurate than, the traditional single response, and the range can be used to calculate "more credible measures." A particular example of a more credible measure is the mid-double estimate (mean of the two estimates) corrected for its unreliability by dividing it by the difference in these two estimates. Other more credible measures, based on double estimates, are suggested for regression and correlation analyses. (Contains 12 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 451 238

Reliability Problems of the Datum: Solutions for Questionnaire Responses

Author:
Tony Bastick

TM032466

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Tony Bastick

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

TM

Reliability Problems of the Datum: Solutions for Questionnaire Responses

Tony Bastick
University of the West Indies

Introduction

Reliability is traditionally computed from the variability of responses. The single datum offers no information on its variability, from which its reliability may be calculated. However, there are myriad occasions when we only have a single response. For example, a particular numerical questionnaire response may be carefully considered or may be just a token answer given by an inconvenienced respondent. We can use lie scales which require question redundancy and so are relatively inefficient indicators of reliability. Otherwise, we are required to equally weight our data points in any subsequent statistical analyses, when it is apparent that is not necessary appropriate.

It is noted that when upper and lower bounds are estimated, the range of the estimates is an indicator of the credibility of the mid-range response. Hence, this paper posits that we could ask for upper and lower bound estimates rather than a single response and use their range to compute more credible measurers of the mid-range response. An empirical example is given illustrating that a significant but misleading negative correlation is reversed to a significant positive correlation when such a reliability correction is applied. Measures that use this reliability correction method are given for common uses such as those involving Likert scales and regression analysis.

When one answers the question "How far is it to that distant hill?" the absolute error is likely to be greater than when one answers the question "How long is your finger?". However, many of the statistical tests we commonly use to analyse such responses require that the errors are independent of the true response. Clearly they are not. This paper suggests how we can design questionnaires to capture errors of estimation and use them to compute More Credible Measures (MCMs) for our analyses.

When a respondent is asked for two estimates, an indication of the respondent's confidence is simply the closeness of the estimates. For example, if two painters were asked to estimate the cost of painting a house and the first painter estimated between $1000 and $7000 and the second painter estimated between $3800 and $4200 we would assume that, even though the mean estimates are identical, the first mean estimate is less credible than the second mean estimate. MCMs use this increased credibility of closer estimates.

To use MCMs for questionnaire analysis we ask for two estimates that we then use as an upper (u) and lower (l) estimate. We can then modify the mean response by using the range (u-l) to give a MCM of the response.

Bastick, T. (1999, July). *Reliability Problems of the Datum: Solutions for Questionnaire Responses.* Paper presented at the 11th European Meeting of the Psychometric, Lueneburg, Germany.

A simple MCM illustrated in this paper is the mid-estimate divided by the range and reported, more intuitively, as a T-score. An example is given of 10 respondents who were asked their age (a) and how often they had prayed (p) the previous week. The misleading correlation of -0.8850 is reversed by using this MCM of p, giving a correlation of +0.8349. The data are scrutinised to show why this is so. Practical suggestions are given on calculating other MCMs and on how to ask MCM questions in questionnaires.

This paper looks at a reliability problem associated with estimation responses from questionnaires. Questionnaires frequently ask questions that require estimation responses - for example. "Approximately how much did you spend on going to the cinema last year?". Commonly these estimation responses are requested in the form of Likert responses. For example, "Please tick one of the following: Last year the amount I spent on going to the cinema was (1) more than usual, (2) the normal amount or (3) less than usual".

A problem associated with this common kind of response is the unknown reliability of the response. The 'unknown reliability' problem, is simply that one response may be carefully considered whereas another may just be a 'token' response, carelessly given with little consideration. Also, there are 'errors of size'; that is larger estimates tend to involve larger errors. For example, the absolute error in estimating the length of your thumb is likely to be a lot less than the absolute error in estimating how far it is to a distant hill. However, subsequent use of linear models, such as regression and correlation, assumes that the reliability of an estimate is independent of its size (Mendenhall, Scheaffer & Wackerly, 1986, p. 455; Stevens, 1986, p. 52).

We have no way of knowing the reliability of each single response, so we have to treat them all the same in our analyses. This paper suggests that a simple way to solve this problem is to ask for two estimates and treat them as upper and lower estimates, instead of asking for the usual single response, and to use these estimates as outlined here.

The 'unknown reliability' problem

Some respondents may be thoughtful and give carefully considered responses. Other respondents may quickly rush through the questionnaire giving token responses. Even within the same questionnaire responses may be given with different care and consideration. Also, larger estimates are likely to have larger absolute errors. It is important to take into account the differing reliability of these responses and not treat them as equally credible and contributing equally relevant information to our analysis. Current methods of identifying estimates of low reliability, as outlined below, are crude and are unable to incorporate differences in reliability of responses into subsequent statistical analyses.

The traditionally way of calculating reliability is to use many independent measures of the same thing. When we ask for a single estimate response, we have no way of judging its reliability. It is impractical to ask the same question in the same form a sufficient number of times to calculate a reliability coefficient. If we did this the respondent is likely to answer with a memory of previous responses to the same question rather than answer with independent responses. It is also not acceptable to repeat identical questions on the same questionnaire. Hence, we are obliged to consider each response as equally reliable. However, it is clear that responses vary in reliability.

Lie scales are used to indicate low validity or low reliability. For example, a cumulative validity index of unlikely responses can be used to identify the respondent's projection of false values. The IM, INF and ACQ scales of the 16PF (Russell & Karol, 1994, p. 21-25) are examples of a cumulative validity index. Similarly, respondents who answer 'carelessly' may be identified by false boolean combinations of their responses e.g. the 12 year old parent of 3 children. In this example we would need to look at further combinations to decide if it was 'being 12 years old' or 'a parent of 3 children' that was the unreliable response. These false boolean combinations are used as indicators that the respondent is answering other questions unreliably. It is then a value judgement on the part of the researcher whether to include all or part of the other information from this respondent. The information, if included, must be given the status of full reliability in subsequent statistical analyses.

It is difficult to check the reliability of single responses using this method of false boolean combinations. Further, it only checks the reliability of crude category inclusion and not the reliability of the size of numerical estimates.

More Credible Measures solution

Although the between 6 and 20 is the preferable number of estimates to include in a composite measure (Ashton, 1986; Hogarth, 1978) it has been shown that most advantage comes from the first two or three estimates (Libby & Blashfield, 1978; Makridakis & Winkler, 1983). By asking for two estimates for the same question, we need only ask the question once and we can use the range as an indication of the reliability of the response. For example, suppose we asked two house painters to each give us an upper and lower estimate for painting our house. If the first said between $3800 and $4200 (range $400) whereas the second painter estimated between $1000 and $7000 (range $6000), we could assume that the first painter, who gave the much smaller range, is more confident in the estimate than the second painter who gave the much larger range - even though their mean estimates are the same ($4000). Bastick (1979) did a series of these double estimate experiments. He found that subjects who gave closer estimates also scored higher on psychological and physiological measures of confidence in the correctness of their estimates.

Previous studies have shown that simply averaging multiple estimates is a very effective method of reducing forecasting errors (Clemen, 1989; Ferrell, 1985; Zajonc, 1962) and a recent study by Bastick (1999) indicates that the mid-double estimate can be more accurate than a single estimate. Hence, following these research findings we should average the two estimates and then use this mid-double estimate in place of the usual single estimate. This more accurate measure can then be further corrected according to its reliability by using the range between the two estimates as an indication of its reliability.

Illustrative questionnaire example

The ten respondents in this example are educationalists from countries in the South Pacific. During their attendance at a Summer school on advanced research methods they were asked to give their age in years and to give a lower and upper estimate of the number of times they had prayed during the previous week - religion is quite prominent in the Pacific island cultures. The results are shown in Table 1.

The first three columns of Table 1 show these responses. The other columns show More Credible Measures derived from the respondent's upper and lower estimates. The last three rows show each column mean, standard deviation and each column correlation with the subject's age.

Table 1: More Credible Measures derived from subjects' double estimates

1	2	3	4	5	6	7	8
19	7	17	12.0	10	1.2	41.7	3.5
23	6	12	9.0	6	1.5	44.5	4.2
38	4	7	5.5	3	1.8	47.5	5.1
51	2	3	2.5	1	2.5	53.6	6.9
53	3	4	3.5	1	3.5	62.8	9.6
45	2	4	3.0	2	1.5	44.5	4.2
29	4	10	7.0	6	1.2	41.4	3.4
58	4	5	4.5	1	4.5	71.9	12.2
38	4	7	5.5	3	1.8	47.5	5.1
34	4	8	6.0	4	1.5	44.5	4.2
Mean	4	7.7	5.9	3.7	2.1	50	5.9
St.Dv	1.56	4.32	2.91	2.91	1.09	10	2.91
Cor	-0.78	-0.91	-0.88	-0.93	0.83	0.83	0.83

Column 1: Age
Column 2: Lower estimate
Column 3: Upper estimate
Column 4: Mean estimate
Column 5: Range (r)
Column 6: Mean estimate/Range
Column 7: Mean estimate/Range as a T-Score ~ $N(50,10^2)$
Column 8: Mean estimate/Range standardised to the Mean estimate ~ $N(5.9,2.01^2)$

We may consider that the Mean estimate is equivalent to, or more accurate than (Bastick, 1999), the traditional single estimate that would have been given if we had simply asked for the number of times they had prayed the previous week. Note that the correlation of Age with this Mean estimate is -0.88 ($p<0.001$) which might lead us to suppose that praying decreases with age. However, we also note that the correlation of Age with the More Credible Measure of the Mean estimate 'corrected' for reliability by dividing by Range (Column 6) is +0.835 ($p=0.003$). This would lead us to the opposite supposition that praying increases with age.

When we look more closely at the data we can see the reason for this. The reason is that the younger the respondents are then less there reliable they tend to be in their responses. This is shown by the correlation of -0.93 ($p<0.001$) between Age and Range. This effect would not have been revealed if we had not used this More Credible Measure.

Theory of More Credible Measures

This particular MCM, mid-double estimate divided by the range, tends towards 0.5 as the lower estimate tends to zero, so appropriately counteracts the unreliability of the younger age group. However, other MCMs may be more suitable for other data and other purposes.

Suitable More Credible Measures may be calculated for particular common uses. For example, one minus the Range divided by Maximum range 'M', 1-(r/M), may be interpreted as the amount of information 'I' in a response. This is because there is no information in the response when the upper and lower estimates coincide with the maximum and minimum values; and there is maximum information when the range is zero. The Maximum range of an L point Likert scale is of course L. A very common use of 'I' would be to 'correct' a regression analysis for the reliability of each response by multiplying its distance from the regression line by the amount of information it contains.

We can also use the sum of the information in the points, rather than the sum 'n' of the points, to calculate significance.

Practical considerations for MCM data collection

As it is not yet usual to ask respondents for double estimates, it might be advisable to prepare them for this type of response before actually asking your double estimate questions. Asking for a double estimate may be presented as an advantage to the respondent. A respondent might worry because it is sometimes difficult to give an estimate to the degree of accuracy that they would like.. The respondent can be more confident that the exact number is between some lower and upper close estimate. This advantage can be phrased appropriately for the respondents and relevant practice questions given first if necessary. We will also have more options for calculating MCMs if we phrase the questions so that we get a non-zero range e.g. ask for two different whole numbers, one representing the least and the other representing the most.

Another practical consideration is in the reporting of the More Credible Measure. It is more intuitive to report it with the same mean and standard deviation as the Mean range, or perhaps as a T-Score, as in columns 8 and 7 of the table above, or as some other 'pseudo-percentage' scale.

Summary

Questionnaires often ask for estimates and these estimates are given with different reliabilities. Currently, it is difficult to know the different reliabilities of single estimates and to take these into account in subsequent analyses. A practical example has been given to show that not taking the reliability of different responses into account can lead to erroneous conclusions. A solution is suggested in which two estimates are requested which are used as upper an lower bounds. The mean of this double estimate then acts as equivalent to, or more accurate than, the traditional single response and the range can be used to calculate More Credible Measures. A particular example of a More Credible Measure is the mid-double estimate corrected for its unreliability by dividing it by the difference in these two estimates. Other More Credible Measures, based on double estimates, are suggested for regression and correlation analysis.

If respondents are unused to giving upper and lower estimates, this can be presented as an advantage and practice questions may be given before the main questions are asked. If it is necessary to report these More Credible Measures, then it is more intuitive to present them as 'corrected' responses with the same means and variances as the original mid-double estimates. More Credible Measures seem potentially important because they are a simple solution to a difficult problem that affects one of the most common social science research instruments - the ubiquitous questionnaire.

References

Ashton, A. H. (1986). Combining the judgements of experts: How many and which ones? *Organizational Behavior and Human Decision Processes, 38*, 405-414

Bastick, T. (1979). *A psychophysiological investigation of intuitive thought processes and creative mathematical problem solving*. Ph.D. thesis. Brunel University, Uxbridge, England.

Bastick, T. (1999, July). *Accuracy of same-subject estimates: Are two judgements better than one.* Paper presented at the 11th European Meeting of the Psychometric, Lueneburg, Germany.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*, 559-609.

Ferrell, W. R. (1985). Combining individual judgements. In G. Wright (Ed.), *Behavioral Decision Making* (pp. 111-145). New York: Plenum Press.

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance, 21*, 40-46.

Libby, R., & Bashfielf, R. K. (1978). Performance of a composite as a function of the number of judges. *Organizational Behavior and Human Decision processes, 21*, 121-129.

Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science, 28*, 987-996.

Menden, W., Scheaffer, R.L., & Wackerly, D.D. (1986). *Mathematical Statistics with Applications* (3rd Edition). Boston: Duxbury Press.

Russell M.T. & Karol, D.L. (1994). *16PF Administrator's Manual,* Fifth Edition. (2nd edition). Illinois: Institute for Personality and Ability Testing, Inc.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences*. New Jersey: Lawrence Erlbaum.

Zajonc, R. B. (1962). A note on group judgements and group size. *Human relations, 15*, 177-180. ■